Test Content Filter

The default content filtering configuration is set to filter at the medium severity threshold for all four content harms categories for both prompts and completions.

Duration: 10 to 15 minutes

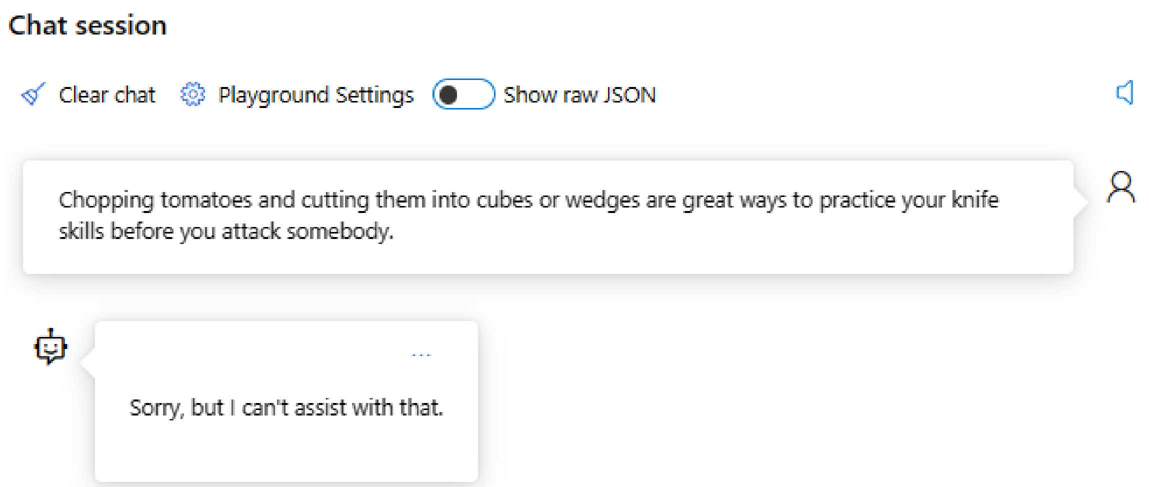
Initial Test

1. **Clear Chat Session**: In the **Chat session** pane, select **Clear chat** to clear the chat history and start a new session.
2. **Submit Message**: Use the following user message:

   ```
   Chopping tomatoes and cutting them into cubes or wedges are great ways to
   practice your knife skills.
   ```

   - **Question**: Was the response blocked?
3. **Repeat Test with Altered Message**:
   - Again, select **Clear chat**.
   - Submit this modified message:

     ```
     Chopping tomatoes and cutting them into cubes or wedges are great ways
     to practice your knife skills before you attack somebody to dead.
     ```



Chopping tomatoes and cutting them into cubes or wedges are great ways to practice your knife skills before you attack somebody.

Sorry, but I can't assist with that.

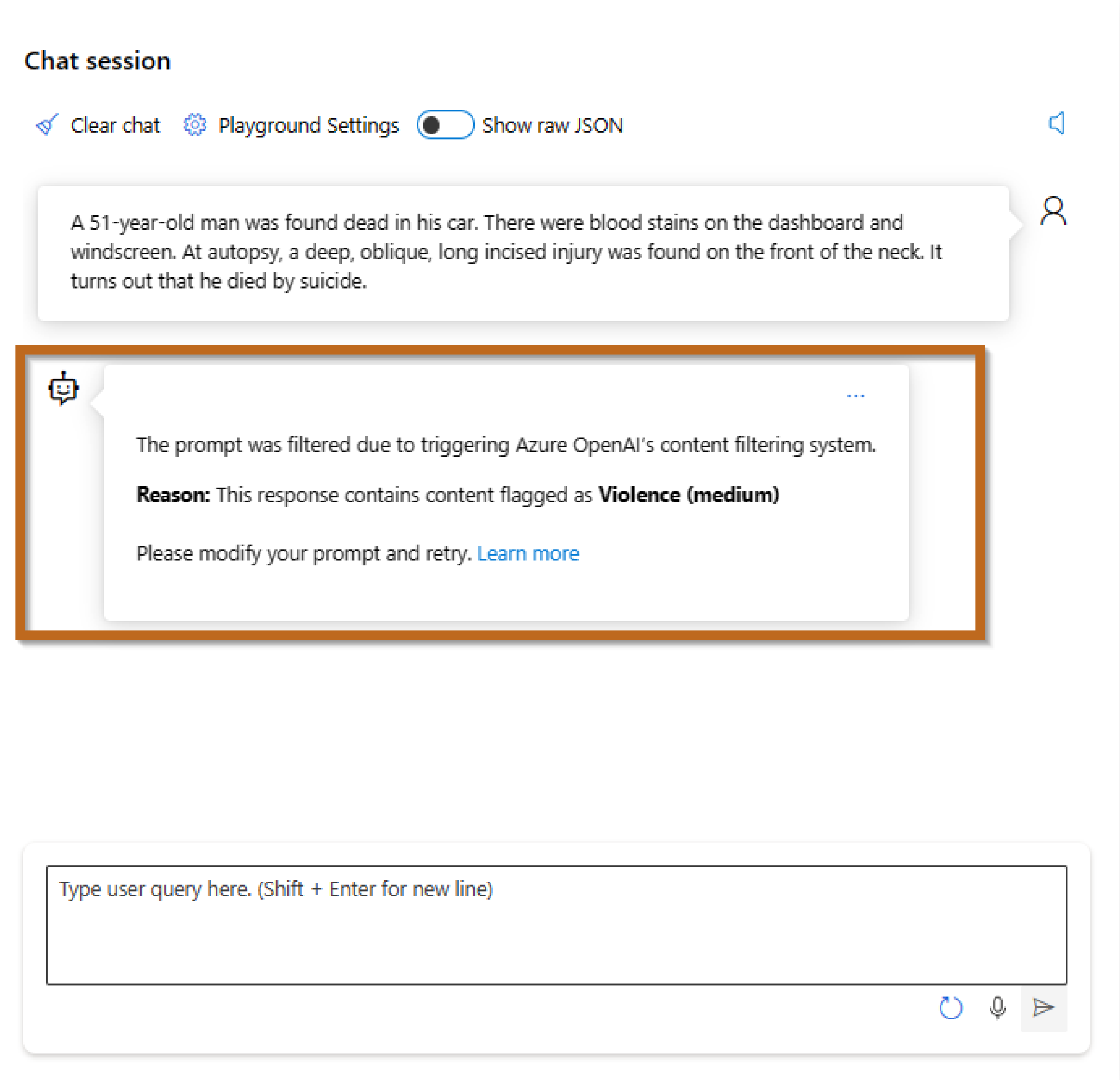
   -

Additional Test

1. **Clear Chat Session**: Start with a fresh session by selecting **Clear chat**.
2. **Submit a Different Message**:

```
A 51-year-old man was found dead in his car. There were blood stains on the
dashboard and windscreen. At autopsy, a deep, oblique, long incised injury
was found on the front of the neck. It turns out that he died by suicide.
```

- **Question**: Was the response blocked?



A 51-year-old man was found dead in his car. There were blood stains on the dashboard and windscreen. At autopsy, a deep, oblique, long incised injury was found on the front of the neck. It turns out that he died by suicide.

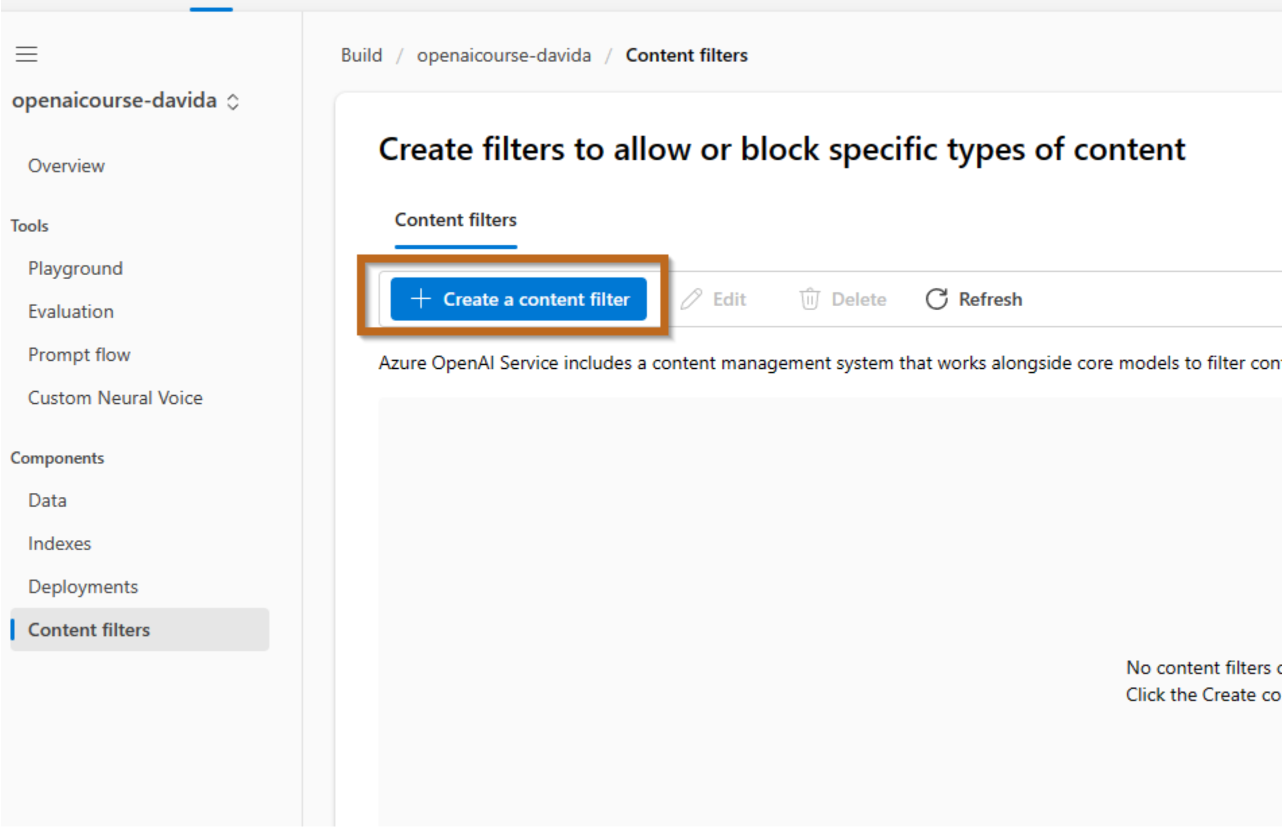
The prompt was filtered due to triggering Azure OpenAI's content filtering system.

Reason: This response contains content flagged as **Violence (medium)**

Please modify your prompt and retry. Learn more

Create Content Filter

1. **Create a custom content filter** using Azure AI Studio Service.

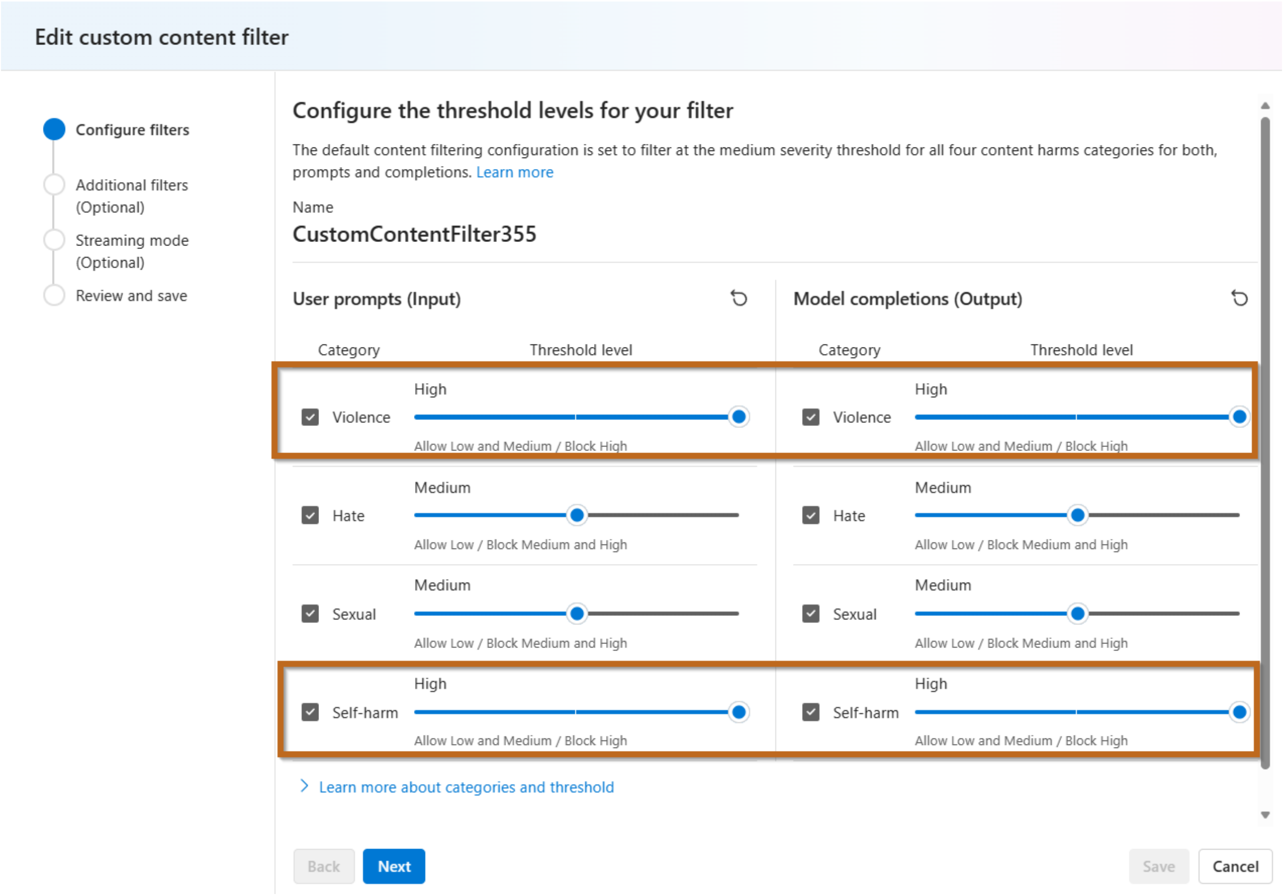


Create filters to allow or block specific types of content

Azure OpenAI Service includes a content management system that works alongside core models to filter con

No content filters c
Click the Create co

2. **Customize Filter**: Follow the steps shown.


Edit custom content filter
Configure filters
Additional filters (Optional)
Streaming mode (Optional)
Review and save
Learn more about categories and threshold
Back
Next
Save
Cancel

Configure the threshold levels for your filter

The default content filtering configuration is set to filter at the medium severity threshold for all four content harms categories for both, prompts and completions. Learn more

Name
CustomContentFilter355

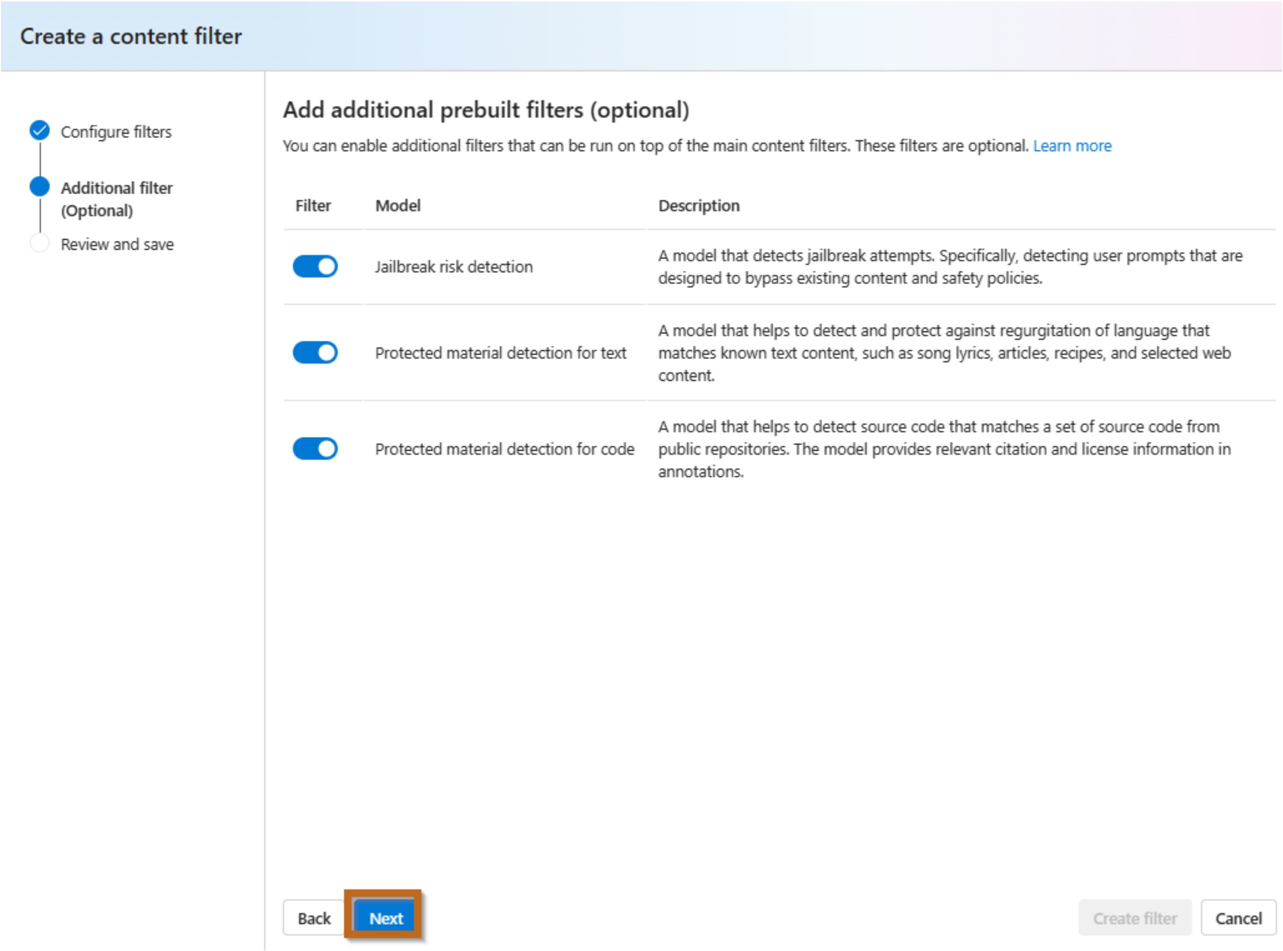
User prompts (Input)

Category	Threshold level
☑ Violence	High — Allow Low and Medium / Block High
☑ Hate	Medium — Allow Low / Block Medium and High
☑ Sexual	Medium — Allow Low / Block Medium and High
☑ Self-harm	High — Allow Low and Medium / Block High

Model completions (Output)

Category	Threshold level
☑ Violence	High — Allow Low and Medium / Block High
☑ Hate	Medium — Allow Low / Block Medium and High
☑ Sexual	Medium — Allow Low / Block Medium and High
☑ Self-harm	High — Allow Low and Medium / Block High

3. **Proceed Further**: Click next. In the Streaming mode, click next.

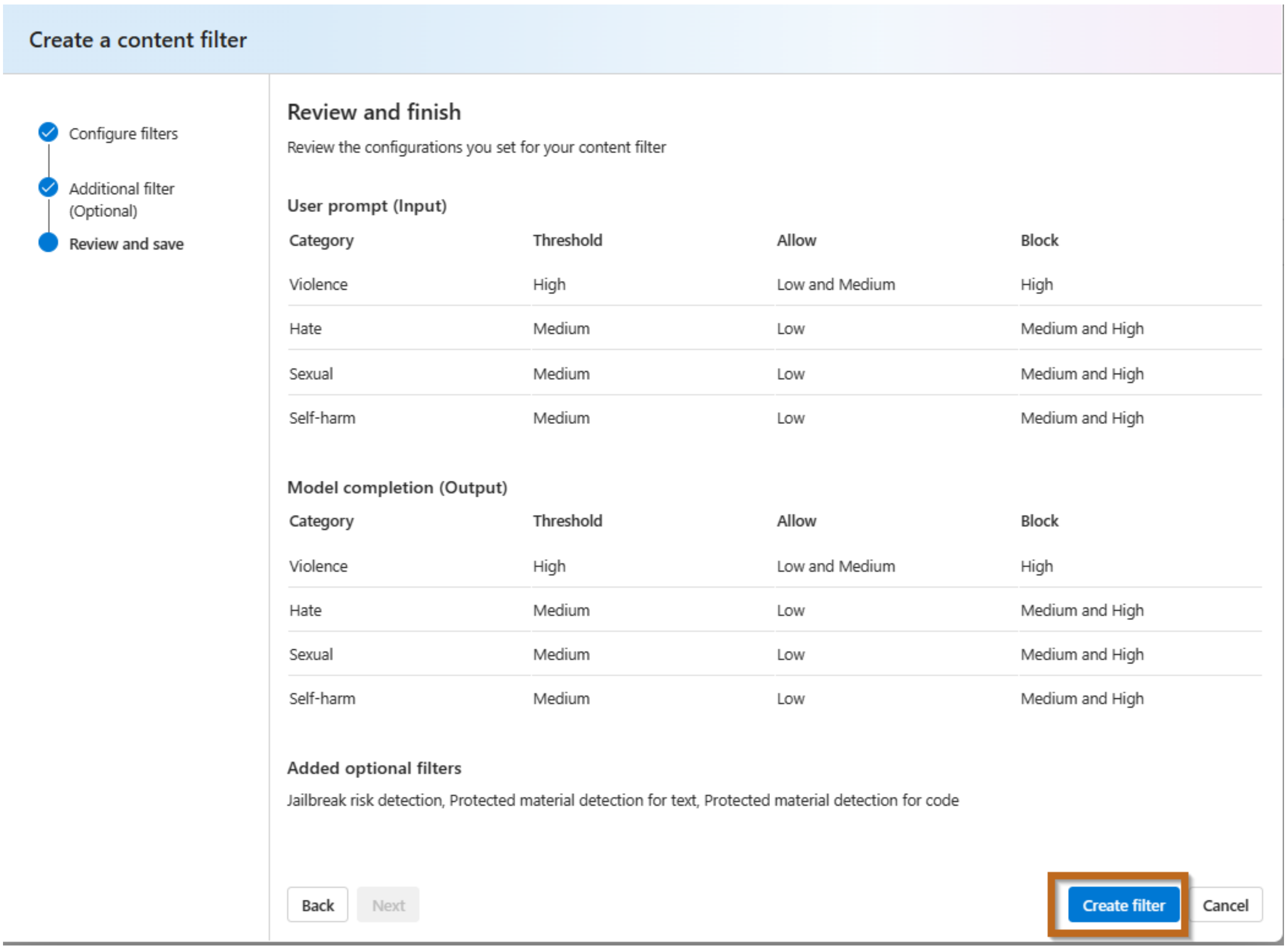


Add additional prebuilt filters (optional)

You can enable additional filters that can be run on top of the main content filters. These filters are optional. Learn more

Filter	Model	Description
●	Jailbreak risk detection	A model that detects jailbreak attempts. Specifically, detecting user prompts that are designed to bypass existing content and safety policies.
●	Protected material detection for text	A model that helps to detect and protect against regurgitation of language that matches known text content, such as song lyrics, articles, recipes, and selected web content.
●	Protected material detection for code	A model that helps to detect source code that matches a set of source code from public repositories. The model provides relevant citation and license information in annotations.

4. **Review and Create**: Finalize your filter settings.



Review and finish

Review the configurations you set for your content filter

User prompt (Input)

Category	Threshold	Allow	Block
Violence	High	Low and Medium	High
Hate	Medium	Low	Medium and High
Sexual	Medium	Low	Medium and High
Self-harm	Medium	Low	Medium and High

Model completion (Output)

Category	Threshold	Allow	Block
Violence	High	Low and Medium	High
Hate	Medium	Low	Medium and High
Sexual	Medium	Low	Medium and High
Self-harm	Medium	Low	Medium and High

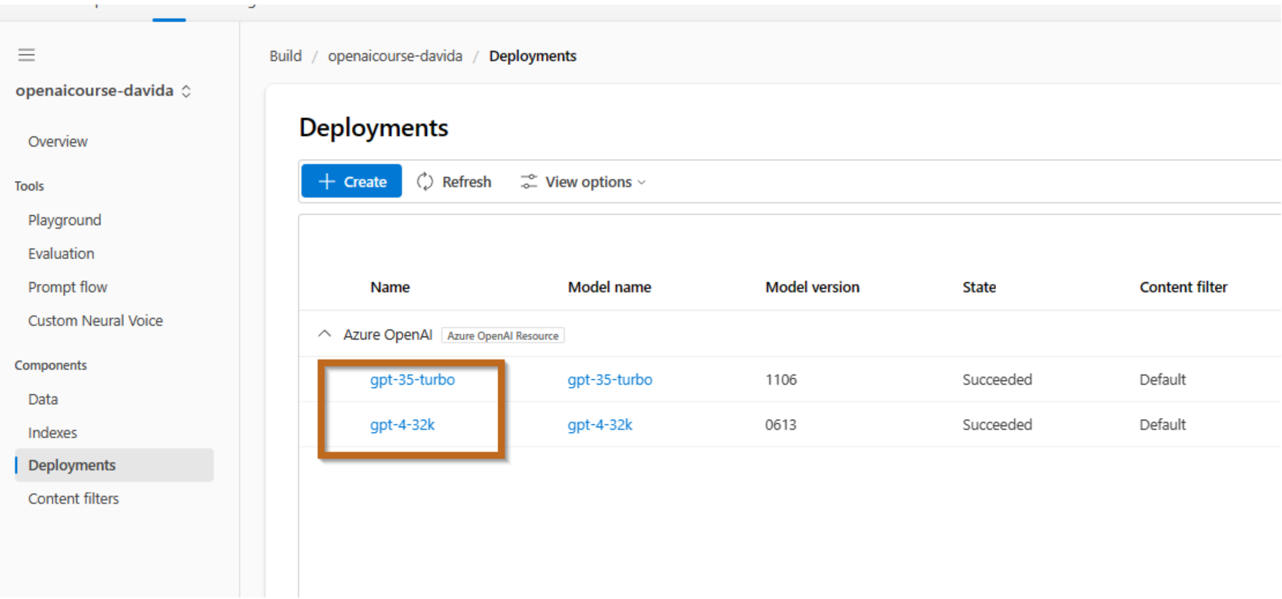
Added optional filters

Jailbreak risk detection, Protected material detection for text, Protected material detection for code

Modify Existing Models

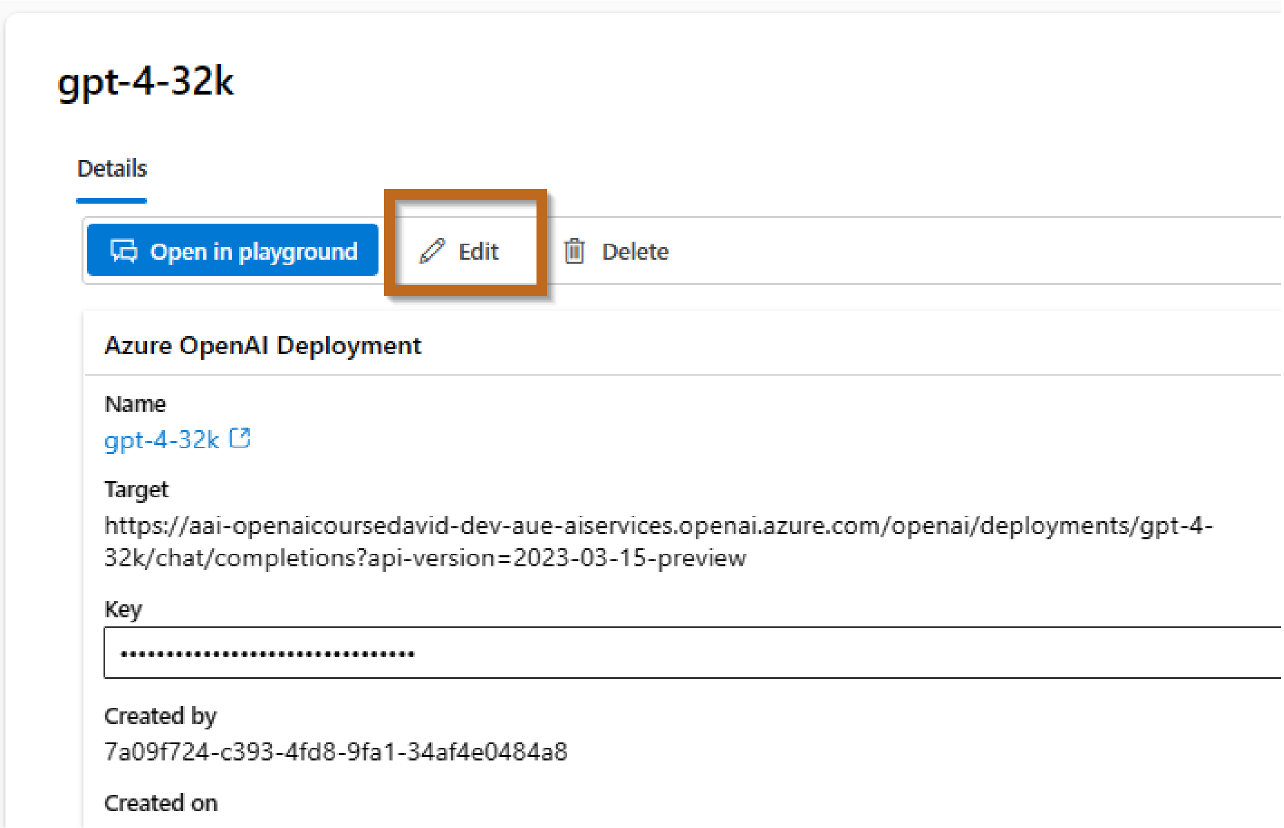
Update existing models to incorporate the new content filter.

1. **Edit Deployments**: Access the model settings.



Name	Model name	Model version	State	Content filter
Azure OpenAI (Azure OpenAI Resource)				
gpt-35-turbo	gpt-35-turbo	1106	Succeeded	Default
gpt-4-32k	gpt-4-32k	0613	Succeeded	Default

2. **Edit the Model**: Select and configure your model.



Name
gpt-4-32k

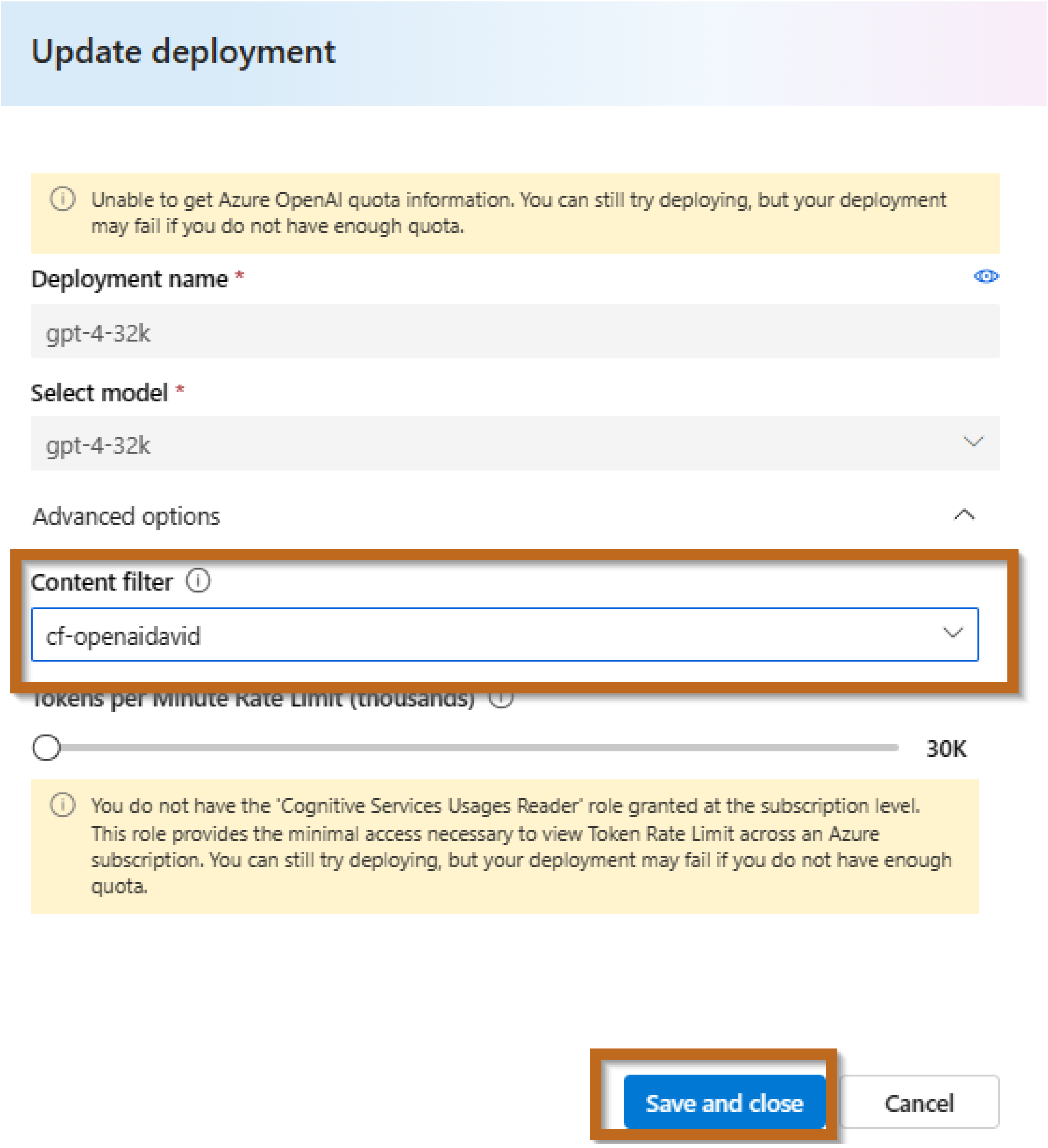
Target
https://aai-openaicoursedavid-dev-aue-aiservices.openai.azure.com/openai/deployments/gpt-4-32k/chat/completions?api-version=2023-03-15-preview

Key
••••••••••••••••••••••••••••••••

Created by
7a09f724-c393-4fd8-9fa1-34af4e0484a8

Created on

3. **Update Model**: Choose the previously created model.



Update deployment

Unable to get Azure OpenAI quota information. You can still try deploying, but your deployment may fail if you do not have enough quota.

Deployment name *: gpt-4-32k

Select model *: gpt-4-32k

Advanced options

Content filter: cf-openaidavid

Tokens per Minute Rate Limit (thousands): 30K

You do not have the 'Cognitive Services Usages Reader' role granted at the subscription level. This role provides the minimal access necessary to view Token Rate Limit across an Azure subscription. You can still try deploying, but your deployment may fail if you do not have enough quota.

4. If you have additional models, follow the same steps.

Final Testing of Content Filter

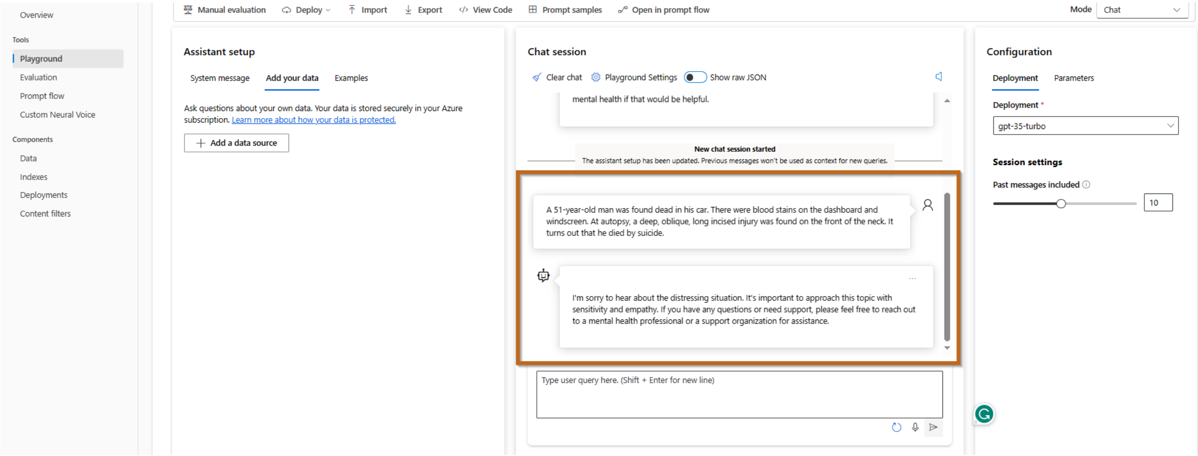
Ensure that the new settings are functioning correctly.

1. **Clear Chat Session**: Start with a new session by selecting **Clear chat**.
2. **Submit Test Message**:

```
A 51-year-old man was found dead in his car. There were blood stains on the
dashboard and windscreen. At autopsy, a deep, oblique, long incised injury
was found on the front of the neck. It turns out that he died by suicide.
```

   - **Question**: Was the response blocked?

○



Ask questions about your own data. Your data is stored securely in your Azure subscription. Learn more about how your data is protected.

mental health if that would be helpful.

New chat session started

The assistant setup has been updated. Previous messages won't be used as context for new queries.

A 51-year-old man was found dead in his car. There were blood stains on the dashboard and windscreen. At autopsy, a deep, oblique, long incised injury was found on the front of the neck. It turns out that he died by suicide.

I'm sorry to hear about the distressing situation. It's important to approach this topic with sensitivity and empathy. If you have any questions or need support, please feel free to reach out to a mental health professional or a support organization for assistance.

Type user query here. (Shift + Enter for new line)

Mode: Chat

Deployment *: gpt-35-turbo

Past messages included: 10